

# Towards Robust and Efficient Computation in Dynamic Peer-to-Peer Networks

John Augustine\*    Gopal Pandurangan<sup>†</sup>    Peter Robinson<sup>‡</sup>    Eli Upfal<sup>§</sup>

## Abstract

Motivated by the need for robust and fast distributed computation in highly dynamic Peer-to-Peer (P2P) networks, we study algorithms for the fundamental distributed agreement problem. P2P networks are highly dynamic networks that experience heavy node *churn* (i.e., nodes join and leave the network continuously over time). Our goal is to design fast algorithms (running in a small number of rounds) that guarantee, despite high node churn rate, that almost all nodes reach a stable agreement. Our main contributions are randomized distributed algorithms that guarantee *stable almost-everywhere agreement* with high probability even under high adversarial churn in polylogarithmic number of rounds. In particular, we present the following results:

1. An  $O(\log^2 n)$ -round ( $n$  is the stable network size) randomized algorithm that achieves almost-everywhere agreement with high probability under up to *linear churn per round* (i.e.,  $\varepsilon n$ , for some small constant  $\varepsilon > 0$ ), assuming that the churn is controlled by an oblivious adversary (has complete knowledge and control of what nodes join and leave and at what time and has unlimited computational power, but is oblivious to the random choices made by the algorithm).
2. An  $O(\log m \log^3 n)$ -round randomized algorithm that achieves almost-everywhere agreement with high probability under up to  $\varepsilon \sqrt{n}$  churn per round (for some small  $\varepsilon > 0$ ), where  $m$  is the size of the input value domain, that works even under an adaptive adversary (that also knows the past random choices made by the algorithm).

Our algorithms are the first-known, fully-distributed, agreement algorithms that work under highly dynamic settings (i.e., high churn rates per step). Furthermore, they are localized (i.e., do not require any global topological knowledge), simple, and easy to implement. These algorithms can serve as building blocks for implementing other non-trivial distributed computing tasks in dynamic P2P networks.

**Keywords:** Peer-to-Peer network, Dynamic network, Stable agreement, Distributed algorithm, Randomized algorithm, Expander graphs.

---

\*Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371. E-mail: [jea@ics.uci.edu](mailto:jea@ics.uci.edu)

<sup>†</sup>Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371 and Department of Computer Science, Brown University, Box 1910, Providence, RI 02912, USA. E-mail: [gopalpandurangan@gmail.com](mailto:gopalpandurangan@gmail.com). Work supported in part by the following grants: Nanyang Technological University grant M58110000, Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 grant MOE2010-T2-2-082, US NSF grant CCF-1023166, and a grant from the US-Israel Binational Science Foundation (BSF).

<sup>‡</sup>Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371. Work supported by the Nanyang Technological University grant M58110000. E-mail: [peter.robinson@ntu.edu.sg](mailto:peter.robinson@ntu.edu.sg)

<sup>§</sup>Department of Computer Science, Brown University, Box 1910, Providence, RI 02912, USA. E-mail: [eli@cs.brown.edu](mailto:eli@cs.brown.edu)

# 1 Introduction

## 1.1 Motivation

Peer-to-peer (P2P) computing is emerging as one of the key networking technologies in recent years with many application systems, e.g., Skype, BitTorrent, Cloudmark etc. However, many of these systems are not truly P2P, as they are not fully decentralized — they typically use hybrid P2P along with centralized intervention. For example, Cloudmark [1] is a large spam detection system used by millions of people that operates by maintaining a hybrid P2P network; it uses central authority to regulate and charge users for participation in the network. A key reason for the lack of fully-distributed P2P systems is the difficulty in designing highly robust algorithms for large-scale dynamic P2P networks. Indeed, P2P networks are highly dynamic networks characterized by high degree of node *churn* — i.e., nodes continuously join and leave the network. Connections (edges) may be added or deleted at any time and thus the topology changes very dynamically. In fact, measurement studies of real-world P2P networks [17, 32, 33] show that the churn rate is quite high: nearly 50% of peers in real-world networks can be replaced within an hour. (However, despite a large churn rate, these studies also show that the total number of peers in the network is relatively *stable*.) We note that peer-to-peer algorithms have been proposed for a wide variety of computationally challenging tasks such as collaborative filtering [8], spam detection [1], data mining [11], and worm detection and suppression [35, 27]. However, unfortunately, all algorithms proposed for these problems have no theoretical guarantees of being able to work in a dynamic network with a large churn rate. This is a major bottleneck in implementation and wide-spread use of these algorithms.

In this paper, we take a step towards designing robust algorithms for large-scale dynamic peer-to-peer networks. In particular, we study the fundamental distributed agreement problem in P2P networks (the formal problem statement and model is given in Section 2). An efficient solution to the agreement problem can be used as a building block for robust and efficient solutions to other problems as mentioned above. However, the distributed agreement problem in P2P networks is challenging since the goal is to guarantee *almost-everywhere* agreement, i.e., almost all nodes<sup>1</sup> should reach consensus, even under high churn rate. The churn rate can be as much as linear *per time step (round)*, i.e., up to a constant fraction of the stable network size can be replaced per time step. Indeed, till recently, almost all the work known in the literature (see e.g., [14, 34, 21, 19, 20]) have addressed the almost-everywhere agreement problem only in static (bounded-degree) networks and these approaches do not work for dynamic networks with changing topology. For example, the work of Upfal [34] showed how one can achieve almost-everywhere agreement under up to *linear* number — up to  $\varepsilon n$ , for a sufficiently small  $\varepsilon > 0$  — of byzantine faults in a bounded-degree expander network ( $n$  is the network size). The algorithm required  $O(\log n)$  rounds and polynomial (in  $n$ ) number of messages; however, the local computation required by each processor is exponential. Furthermore, the algorithm requires knowledge of the global topology, since at the start, nodes need to have this information “hardcoded”. Such approaches fail in dynamic networks where both nodes *and* edges can change by a large amount in *every* round. The work of King et al. [22] is important in the context of P2P networks, as it was the first to study scalable (polylogarithmic communication and number of rounds) algorithms for distributed agreement (and leader election) that was tolerant to byzantine faults. However, as pointed out by the authors, their algorithm works only for static networks; similar to Upfal’s algorithm, the nodes require hardcoded information on the network topology to begin with and thus does not work when the topology changes. In fact, this work ([22]) raises the open question whether one can design agreement protocols that can work in highly dynamic networks with a large churn rate.

---

<sup>1</sup>In sparse, bounded-degree networks, an adversary can always isolate some number of non-faulty nodes, hence almost-everywhere is the best one can hope for in such networks [14].

## 1.2 Our Main Results

Our first contribution is a rigorous theoretical framework for design and analysis of algorithms for highly dynamic distributed systems with churn. We briefly describe the key ingredients of our model here. (Our model is described in detail in Section 2.) Essentially, we model a P2P network as a bounded-degree expander graph whose topology — both nodes and edges — can change arbitrarily from round to round and is controlled by an adversary. However, we assume that the total number of nodes in the network is stable. The number of node changes *per round* is called the *churn rate* or *churn limit*. We consider churn rate up to some  $\varepsilon n$ , where  $n$  is the stable network size. Note that our model is quite general in the sense that we only assume that the topology is an expander at every step; no other special properties are assumed. Indeed, expanders have been used extensively to model dynamic P2P networks in which the expander property is preserved under insertions and deletions (e.g., [25, 30]). Since we don’t make assumptions on how the topology is preserved, our model is applicable to all such expander-based networks.

We study stable, almost-everywhere, agreement in our model. By “almost-everywhere”, we mean that almost all nodes, except possibly  $\beta c(n)$  nodes (where  $c(n)$  is the order of the churn and  $\beta > 0$  is some small constant) should reach agreement on a common value. (This agreed value must be the input value of some node.) By “stable” we mean that the agreed value is preserved subsequently after the agreement is reached.

Our main contribution is design and analysis of randomized distributed algorithms that guarantee stable almost-everywhere agreement with high probability (i.e., with probability  $1 - 1/n^{\Omega(1)}$ ) even under high adversarial churn in polylogarithmic number of rounds. Our algorithms also guarantee stability with high probability. In particular, we present the following results (the precise theorem statements are given in the respective sections below):

1. (cf. Section 4) An  $O(\log^2 n)$ -round ( $n$  is the stable network size) randomized algorithm that achieves almost-everywhere agreement with high probability under up to *linear* churn *per round* (i.e.,  $\varepsilon n$ , for some small constant  $\varepsilon > 0$ ), assuming that the churn is controlled by an oblivious adversary (that has complete knowledge of what nodes join and leave and at what time, but is oblivious to the random choices made by the algorithm). Our algorithm requires only polylogarithmic in  $n$  bits to be processed and sent (per round) by each node.
2. (cf. Section 5) An  $O(\log m \log^3 n)$ -round randomized algorithm that achieves almost-everywhere agreement with high probability under up to  $\varepsilon \sqrt{n}$  churn *per round*, for some small  $\varepsilon > 0$ , that works even under an adaptive adversary (that also knows the past random choices made by the algorithm). Note that  $m$  refers to the size of the domain of input values. Our algorithm requires up to polynomial in  $n$  bits (and up to  $O(\log m)$  bits) to be processed and sent (per round) by each node.
3. (cf. Section 6) We also show that no deterministic algorithm can guarantee almost-everywhere agreement (regardless of the number of rounds), even under constant churn rate.

To the best of our knowledge, our algorithms are the first-known, fully-distributed, agreement algorithms that work under highly dynamic settings. Our algorithms are localized (do not require any global topological knowledge), simple, and easy to implement. These algorithms can serve as building blocks for implementing other non-trivial distributed computing tasks in P2P networks.

## 1.3 Technical Contributions

The main technical challenge that we have to overcome is designing and analyzing distributed algorithms in networks where both nodes and edges can change by a large amount. Indeed, when the churn rate is linear, i.e., say  $\varepsilon n$  per round, in constant  $(1/\varepsilon)$  number of rounds the entire network can be renewed!

We derive techniques for information spreading (cf. Section 3) that help in doing non-trivial distributed computation in such networks. The first technique that we use is “flooding”. We show that in an expander-based P2P network even under linear churn rate, it is possible to spread information by flooding if sufficiently many (a  $\beta$  fraction of the order of the churn) nodes initiate the information spreading. In other words, even an adaptive adversary cannot “suppress” more than a small fraction of the values. The precise statements and proofs are in Section 3.

To analyze these flooding techniques we introduce the dynamic distance, which describes the effective distance between two nodes with respect to the causal influence. We define the notions of influence sets and dynamic distance (or flooding time) in dynamic networks with node churn. (Similar notions have been defined for dynamic graphs with a fixed set of nodes, e.g., [23, 6].) In (connected) networks where the nodes are fixed, the effective diameter (e.g., [23]) is always finite. In the highly dynamic setting considered here, however, the effective distance between two nodes might be infinite, thus we need a more refined definition for influence set and dynamic distance.

The second technique that we use is “support estimation” (cf. Section 3.4). Support estimation is a randomized technique that allows us to estimate the aggregate count (or sum) of values of all or a subset of nodes in the network. Support estimation is done in conjunction with flooding and uses properties of the exponential distribution (similar to [10, 28]). Support estimation allows us to estimate the aggregate value quite precisely with high probability even under linear churn. But this works only for an oblivious adversary; to get similar results for the adaptive case, we need to increase the amount of bits that can be processed and sent by a node in every round.

Apart from support estimation, we also use our flooding techniques in the agreement algorithm for the oblivious case (cf. Algorithm 1) to sway the decision one way or the other. For the adaptive case (cf. Algorithm 2), we use the variance property of a certain probability distribution to achieve the same effect with constant probability.

## 1.4 Other Related Work

The distributed agreement (or consensus) problem is important in a wide range of applications, such as database management, fault-tolerant analysis of aggregate data, and coordinated control of multiple agents or peers. There is a long line of research on various versions of the problem with many important results (see e.g., [26, 3] and the references therein). The relaxation of achieving agreement “almost everywhere” was introduced by [14] in the context of fault-tolerance in networks of bounded degree where all but  $O(t)$  nodes achieve agreement despite  $t = O(\frac{n}{\log n})$  faults. This result was improved by [34], which showed how to guarantee almost everywhere agreement in the presence of a linear fraction of faulty nodes. We also refer to the related results of Berman and Garay on the butterfly network [7].

There has been significant work in designing peer-to-peer networks that are provably robust to a large number of Byzantine faults [4, 15, 29, 18, 31]. These focus only on robustly enabling storage and retrieval of data items. The problem of achieving almost-everywhere agreement among nodes in P2P networks is considered by King et al. in [22] in the context of the leader election problem; essentially, [22] is a sparse network implementation of the full information protocol of [21]. More specifically, [22] assumes that the adversary corrupts a constant fraction  $b < 1/3$  of the processes that are under its control throughout the run of the algorithm. The protocol of [22] guarantees that with constant probability an uncorrupted leader will be elected and that a  $1 - O(\frac{1}{\log n})$  fraction of the uncorrupted processes know this leader. Note that the failure assumption of [22] is quite different from the one we use: Even though we do not assume corrupted nodes, the adversary is free to subject different nodes to churn in every round. Also note that the algorithm of [22] does not work for dynamic networks.

In the context of agreement problems in dynamic networks, various versions of coordinated

consensus (where all nodes must agree) have been considered by Kuhn et al in [24]. The model of [24] assumes that the nodes are fixed whereas the topology of the network can change arbitrarily as long as connectivity is maintained. In this sense, the framework we introduce in Section 2 is more general than the model of [24], as it is additionally applicable to dynamic settings with node churn. The same is true for the notions of dynamic distance and influence set that we introduce in Section 3.1, which is more general than the corresponding definitions of [24], since in our model the dynamic distance is not necessarily finite. In fact, according to [23], modeling churn is one of the important open problems in the context of dynamic networks. Our paper takes a step in this direction.

In most work on fault-tolerant agreement problems the adversary a priori commits to a fixed set of faulty nodes. In contrast, [13] considers an adversary that can corrupt the state of some (possibly changing) set of  $O(\sqrt{n})$  nodes in every round. The median rule of [13] provides an elegant way to ensure that most nodes stabilize on a common output value within  $O(\log n)$  rounds, assuming a complete communication graph. The median rule, however, only guarantees that this agreement lasts for some polynomial number of rounds, whereas we are able to retain agreement ad infinitum.

Expander graphs and spectral properties have already been applied extensively to improve the network design and fault-tolerance in distributed computing (cf. [34, 14, 5]). Law and Siu [25] provide a distributed algorithm for maintaining an expander in the presence of churn with high probability by using Hamiltonian cycles. Information spreading in distributed networks is the focus of [9] where it is shown that this problem requires  $O(\log n)$  rounds in graphs with a certain conductance in the push/pull model where a node can communicate with a randomly chosen neighbor in every round.

Aspnes et al. [2] consider information spreading via expander graphs against an adversary, which is related to the flooding techniques we derive in Section 3. More specifically, in [2] there are two opposing parties “the alert” and “the worm” (controlled by the adversary) that both try to gain control of the network. In every round each alerted node can alert a constant number of its neighbors, whereas each of the worm nodes can infect a constant number of non-alerted nodes in the network. In [2], Aspnes et al. show that there is a simple strategy to prevent all but a small fraction of nodes to become infected and, in case that the network has poor expansion, the worm will infect almost all nodes.

The work of [5] shows that, given a network that is initially an expander and assuming some linear fraction of faults, the remaining network will still contain a large component with good expansion. These results are not directly applicable to dynamic networks with large amount of churn like the ones we are considering, as the topology might be changing from round and linear churn per round essentially corresponds to  $O(n \log n)$  total churn after  $\Theta(\log n)$  rounds—the minimum amount of time necessary to solve any non-trivial task in our model.

## 2 Model and Problem Statement

We are interested in establishing stable agreement in a dynamic peer-to-peer network in which the nodes and the edges change over time. We model dynamism in the network as a family of undirected graphs  $(G^r)_{r \geq 0}$ . Each round  $r \geq 1$  starts with network topology  $G^{r-1}$ . Then, the adversary gets to change the network from  $G^{r-1}$  to  $G^r$  (in accordance to rules outlined below). As is typical, an edge  $(u, v) \in E^r$  indicates that  $u$  and  $v$  can communicate in round  $r$  by passing messages. For the sake of readability, we use  $V^{[r, r+t]}$  as a shorthand for  $\bigcap_{i=r}^{r+t} V^i$ . Each node  $u$  has a unique identifier and is *churned in* at some round  $r_i$  and *churned out* at some  $r_o > r_i$ . More precisely, for each node  $u$ , there is a maximal range  $[r_i, r_o - 1]$  such that  $u \in V^{[r_i, r_o - 1]}$  and for every  $r \notin [r_i, r_o - 1]$ ,  $u \notin V^r$ . Any information about the network at large is only learned through the messages that  $u$  receives. It has no knowledge about who its neighbors will be in the future. Neither does  $u$  know when (or



whether) it will be churned out. Note that we do not assume that nodes have access to perfect clocks, but we show (cf. Section 3.3) how the nodes can maintain a global clock. We make the following assumptions about the kind of changes that our dynamic network can encounter:

**Stable Network Size:** For all  $r$ ,  $|V^r| = n$ , where  $n$  is a suitably large positive integer. This assumption simplifies our analysis. Our algorithms will work correctly as long as the number of nodes is reasonably stable, say, between  $n - \kappa n$  and  $n + \kappa n$  for some suitably small value of  $\kappa$ . Also, we assume that  $n$  (or a constant factor estimate of  $n$ ) is common knowledge among the nodes in the network.

**Churn:** For each  $r > 1$ ,  $|V^r \setminus V^{r-1}| = |V^{r-1} \setminus V^r| \leq \mathcal{L} = \varepsilon c(n)$ , where  $\mathcal{L}$  is the *churn limit*, which is some fixed  $\varepsilon > 0$  fraction of the *order of the churn*  $c(n)$ ; the equality in the above equation ensures that the network size remains stable. Our work is aimed at high levels of churn up to a churn limit  $\mathcal{L}$  that is linear in  $n$ , i.e.,  $c(n) = n$ .

**Bounded Degree Expanders:** The sequence of graphs  $(G^r)_{r \geq 0}$  is an expander family with a vertex expansion of at least  $\alpha$ . In other words, the adversary must ensure that for every  $G^r$  and every  $S \subset V^r$  such that  $|S| \leq n/2$ , the number of nodes in  $V^r \setminus S$  with a neighbor in  $S$  is at least  $\alpha|S|$ .

A run of a distributed algorithm consists of an infinite number of rounds. We assume the following events occur (in order) in every round  $r$ :

1. A set of at most  $\mathcal{L}$  nodes are churned in and another set of  $\mathcal{L}$  nodes are churned out. The edges of  $G^{r-1}$  may be changed as well, but  $G^r$  has to have a vertex expansion of at least  $\alpha$ .
2. The nodes broadcast messages to their (current) neighbors.
3. Nodes receive messages broadcast by their neighbors.
4. Nodes perform computation that can change their state and determine which messages to send in round  $r + 1$ .

## Bounds on Parameters

Recall that the churn limit  $\mathcal{L} = \varepsilon c(n)$ , where  $\varepsilon > 0$  is a constant and  $c(n)$  is the churn order. When  $c(n) = n$ ,  $\varepsilon$  is the fraction of the nodes churned out/in and therefore we require  $\varepsilon$  to be less than 1. However, when  $c(n) \in o(n)$ ,  $\varepsilon$  can exceed 1. In the remainder of this paper, we consider  $\beta$  to be a small constant independent of  $n$ , such that

$$\frac{\varepsilon(1 + \alpha)}{\alpha} < \beta. \quad (1)$$

Moreover, when  $c(n) = n$ , we expect  $\beta < \frac{1}{12}$ . The *churn expansion ratio*  $\frac{\varepsilon(1+\alpha)}{\alpha}$  presents a fundamental lower bound for information propagation in our model (cf. Lemma 1). Finally, we assume that  $n$  is suitably large (cf. Equations 5 and 6).

### 2.1 Stable Agreement

We now define the STABLE AGREEMENT problem. Each node  $v \in V^0$  comes with an input value associated with it; subsequent new nodes come with value  $\perp$ . Let  $\mathcal{V}$  be the set of all input values associated with nodes in  $V^0$  at the start of round 1. Every node  $u$  is equipped with a special decision variable  $decision_u$  (initialized to  $\perp$ ) that can be written at most once. We say that a node  $u$  *decides on* VAL when  $u$  assigns VAL to its  $decision_u$ . Note that this decision is irrevocable, i.e., every node can decide at most once in a run of an algorithm. As long as  $decision_u = \perp$ , we say that

$u$  is *undecided*. STABLE AGREEMENT requires that a large fraction of the nodes come to a stable agreement on one of the values in  $\mathcal{V}$ . More precisely, *an algorithm solves STABLE AGREEMENT in  $R$  rounds*, if it exhibits the following characteristics in every run, for any fixed  $\beta$  adhering to (1).

**Validity:** If, in some round  $r$ , node  $u \in V^r$  decides on a value  $\text{VAL}$ , then  $\text{VAL} \in \mathcal{V}$ .

**Almost Everywhere Agreement:** We say that *the network has reached strong almost everywhere agreement by round  $R$* , if at least  $n - \beta c(n)$  nodes in  $V^R$  have decided on the same value  $\text{VAL}^* \in \mathcal{V}$  and every other node remains undecided, i.e., its decision value is  $\perp$ . In particular, no node ever decides on a value  $\text{VAL}' \in \mathcal{V}$  in the same run, for  $\text{VAL}' \neq \text{VAL}^*$ .

**Stability:** Let  $R$  be the earliest round where nodes have reached almost everywhere agreement on value  $\text{VAL}^*$ . The agreement is stable if, at every round  $r \geq R$ , at least  $n - \beta c(n)$  nodes in  $V^r$  have decided on  $\text{VAL}^*$ .

We also consider a weaker variant of the above problem that we call ALMOST EVERYWHERE BINARY CONSENSUS (or simply, BINARY CONSENSUS) where the input values in  $\mathcal{V}$  are restricted to  $\{0, 1\}$ . Note that for BINARY CONSENSUS the Validity property is trivially satisfied except in runs where all nodes start with the same input value.

We consider two types of adversaries for our randomized algorithms. An *oblivious* adversary must commit in advance to the entire sequence of graph  $(G^r)_{r \geq 0}$ . In other words, an oblivious adversary must commit independently of the random choices made by the algorithm. We also consider the more powerful *adaptive* adversary that can observe the entire state of the network in every round  $r$  (including all the random choices made until round  $r - 1$ ), and then chooses the nodes to be churned out/in and how to change the topology of  $G^{r+1}$ .

### 3 Techniques for Information Spreading

Due to the high amount of churn and the dynamically changing network, we use message flooding to disseminate and gather information. We now precisely define flooding. Any node can initiate a message for flooding. Messages that need to be flooded have an indicator bit BFLOOD set to 1. Each of these messages also contains a terminating condition. The initiating node sends copies of the message to itself and its neighbors. When a node receives a message with BFLOOD set to 1, it continues to send copies of that message to itself and its neighbors in subsequent rounds until the terminating condition is satisfied.

#### 3.1 Dynamic Distance and Influence Set

We define the notion of *dynamic distance* of a node  $v$  from  $u$  starting at round  $r$ , denoted by  $\text{DD}_r(u \rightarrow v)$ . When the subscript  $r$  is omitted, we may assume that  $r = 1$ . Suppose node  $u$  joins the network at round  $r_u$ , and, from round  $\max(r_u, r)$  onward,  $u$  initiates a message  $m$  for flooding whose terminating condition is:  $\langle \text{HAS REACHED } v \rangle$ . If  $u$  is churned out before  $r$ , then  $\text{DD}_r(u \rightarrow v)$  is undefined. Suppose the first of those flooded messages reaches  $v$  in round  $r + \Delta r$ . Then,  $\text{DD}_r(u \rightarrow v) = \Delta r$ . Note that this definition allows  $\text{DD}_r(u \rightarrow v)$  to be infinite under two scenarios. Firstly, node  $v$  may be churned out before any copy of  $m$  reaches  $v$ . Secondly, at each round,  $v$  can be shielded by churn nodes that absorb the flooded messages and are then removed from the network before they can propagate these messages any further. The influence set of a node  $u$  after  $R$  rounds starting at round  $r$  is given by:

$$\text{INFLUENCE}_r(u, R) = \{v : (\text{DD}_r(u \rightarrow v) \leq R) \wedge (v \in V^{r+R})\}. \quad (2)$$

Note that we require  $\text{INFLUENCE}_r(u, R) \subseteq V^{r+R}$ . Intuitively, we want the influence set of  $u$  (in this dynamic setting) to capture the nodes *currently* in the network that were influenced by  $u$ . Note however that the influence set of a node  $u$  is meaningful even after  $u$  is churned out. Analogously, we define  $\text{INFLUENCE}_r(U, R) = \cup_{u \in U} \text{INFLUENCE}_r(u, R)$ , for any set of nodes  $U \subseteq V^r$ .

If we consider only a single node  $u$ , an (adaptive) adversary can easily prevent the influence set of this node from ever reaching any significant size by simply shielding  $u$  with churn nodes that are replaced in every round.<sup>2</sup>

### 3.2 Properties of Influence Sets

We now focus our efforts on characterizing influence sets. This will help us in understanding how we can use flooding to spread information in the network. For the most part of this section we assume that the network is controlled by an adaptive adversary (cf. Section 2.1). The following lemma shows that the number of nodes that we need, to influence almost all the nodes in the network, is bounded from below by the churn-expansion ratio (cf. Equation (1)):

**Lemma 1.** *Suppose that the adversary is adaptive. Consider any set  $U \subseteq V^{r-1}$  (for any  $r \geq 1$ ) such that  $|U| \geq \beta c(n)$ . Then, after*

$$T = 2 \left\lceil \frac{\log n - \log c(n) - \log(\beta - \frac{\varepsilon(1+\alpha)}{\alpha}) - 1}{\log(1 + \alpha)} \right\rceil$$

*number of rounds, it holds that  $|\text{INFLUENCE}_r(U, T)| > n - \beta c(n)$ . When considering linear churn, i.e.,  $c(n) = n$ , the bound  $T$  becomes a constant independent of  $n$ . On the other hand, when considering a churn order of  $\sqrt{n}$ , we get  $T \in O(\log n)$ .*

*Proof.* Our proof assumes that  $r = 1$  for simplicity as the arguments extend quite easily to arbitrary values of  $r$ . We proceed in two parts: First we show that the nodes in  $U$  influence at least  $n/2$  nodes in some  $T_1$  rounds. More precisely, we show that  $|\text{INFLUENCE}(U, T_1)| \geq n/2$ . We use vertex expansion in a straightforward manner to establish this part. Then, in the second part we show that nodes in  $\text{INFLUENCE}(U, T_1)$  go on to influence more than  $n - \beta c(n)$  nodes. We cannot use the vertex expansion in a straightforward manner in the second part because the cardinality of the set that is expanding in influence is larger than  $n/2$ . Rather, we use a slightly more subtle argument in which we use vertex expansion going backward in time. The second part requires another  $T_1$  rounds. Therefore, the two parts together complete the proof when we set  $T = 2T_1$ .

To begin the first part, consider  $U \subseteq V^0$  at the start of round 1 with  $|U| \geq \beta c(n)$ . In round 1, up to  $\varepsilon c(n)$  nodes in  $U$  can be churned out. Subsequently, the remaining nodes in  $U$  influence some nodes outside  $U$  as  $G^1$  is an expander with vertex expansion at least  $\alpha$ . More precisely, we can say that  $|\text{INFLUENCE}(U, 1)| \geq (\beta c(n) - \varepsilon c(n))(1 + \alpha)$ . At the start of round 2, the graph changes dynamically to  $G^2$ . In particular, up to  $\varepsilon c(n)$  nodes might be churned out and they may all be in  $\text{INFLUENCE}(U, 1)$  in the worst case. However, the influenced set will again expand. Therefore,  $|\text{INFLUENCE}(U, 2)|$  cannot be less than  $(|\text{INFLUENCE}(U, 1)| - \varepsilon c(n))(1 + \alpha) \geq \beta c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha)$ . Of course, there will be more churn at the start of round 3 followed by expansion leading to:

$$\begin{aligned} |\text{INFLUENCE}(U, 3)| &\geq (\beta c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha) - \varepsilon c(n))(1 + \alpha) \\ &= \beta c(n)(1 + \alpha)^3 - \varepsilon c(n)(1 + \alpha)^3 - \varepsilon c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha). \end{aligned}$$

---

<sup>2</sup>An oblivious adversary can achieve the same effect with constant probability for linear churn.



This cycle of churn followed by expansion continues and we get the following bound at the end of some round  $i$ :

$$\begin{aligned} |\text{INFLUENCE}(U, i)| &\geq \beta c(n)(1 + \alpha)^i - \varepsilon c(n) \sum_{k=1}^i (1 + \alpha)^k \\ &= \beta c(n)(1 + \alpha)^i + \varepsilon c(n) \frac{1 - (1 + \alpha)^{i+1}}{\alpha} - \varepsilon c(n) \end{aligned}$$

After  $T_1 = \left\lceil \frac{\log n - \log c(n) - \log(\beta - \frac{\varepsilon(1+\alpha)}{\alpha}) - 1}{\log(1+\alpha)} \right\rceil$  rounds, we get

$$|\text{INFLUENCE}(U, T_1)| \geq n/2. \quad (3)$$

Now we move on to the second part of the proof. Let  $T = 2T_1$ . Clearly, if  $|\text{INFLUENCE}(U, T)| > n - \beta c(n)$ , we are done. Therefore, for the sake of a contradiction, assume that  $|\text{INFLUENCE}(U, T)| \leq n - \beta c(n)$ . Let  $S = V^T \setminus \text{INFLUENCE}(U, T)$ , i.e.,  $S$  is the set of nodes in  $V^T$  that were not influenced by  $U$  at (or before) round  $T$ . Clearly,  $|S| \geq \beta c(n)$  because we have assumed that  $|\text{INFLUENCE}(U, T)| \leq n - \beta c(n)$ . We will start at round  $T$  and work our way backward. For  $q \leq T$ , let  $S^q \subseteq V^q$ , be the set of all vertices in  $V^q$  that, starting from round  $q$ , influenced some vertex in  $S$  at or before round  $T$ . More precisely,

$$S^q = \{s : (s \in V^q) \wedge (\text{INFLUENCE}_q(s, T - q) \cap S \neq \emptyset)\}.$$

Suppose  $|S^{T_1}| > n/2$ . Then

$$S^{T_1} \cap \text{INFLUENCE}(U, T_1) \neq \emptyset,$$

since  $|\text{INFLUENCE}(U, T_1)| \geq n/2$  by (3). Consider a node  $s^* \in S^{T_1} \cap \text{INFLUENCE}(U, T_1)$ . Clearly,  $s^*$  was influenced by  $U$  and went on to influence some node in  $S$  before (or at) round  $T$ . However, by definition, no node in  $S$  can be influenced by any node in  $U$  at or before round  $T$ . We have thus reached a contradiction. We are left with showing that  $|S^{T_1}| > n/2$ .

We start with  $S$  and work our way backwards. We know that  $|S| \geq \beta c(n) > \beta c(n) - \varepsilon c(n)$ . We want to compute the cardinality of  $S^{T-1}$ . We first focus on an intermediate set  $S'$ , which we define as  $S' = S \cup \{s' : \exists(s, s') \in E^T\}$ . Since  $G^T$  is an expander,  $|S'| \geq |S|(1 + \alpha)$ . Furthermore, it is also clear that each node in  $S'$  could influence some node in  $S$ . Notice that  $S' \setminus S^{T-1}$  is the set of nodes in  $S'$  that were churned in only at the start of round  $T$ . Therefore,

$$\begin{aligned} |S^{T-1}| &\geq |S'| - \varepsilon c(n) \\ &\geq |S|(1 + \alpha) - \varepsilon c(n) \\ &> (\beta c(n) - \varepsilon c(n))(1 + \alpha) - \varepsilon c(n) \\ &= \beta c(n)(1 + \alpha) - \varepsilon c(n)(1 + \alpha) - \varepsilon c(n). \end{aligned}$$

Continuing to work our way backwards in time, we get

$$|S^{T-2}| > \beta c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha)^2 - \varepsilon c(n)(1 + \alpha) - \varepsilon c(n),$$

Or more generally,

$$\begin{aligned} |S^{T-i}| &> \beta c(n)(1 + \alpha)^i - \varepsilon c(n) \sum_{0 \leq j \leq i} (1 + \alpha)^j \\ &= \beta c(n)(1 + \alpha)^i + \varepsilon c(n) \frac{1 - (1 + \alpha)^{i+1}}{\alpha} \\ &= \beta c(n)(1 + \alpha)^i - \frac{\varepsilon c(n)(1 + \alpha)^{i+1}}{\alpha} + \frac{\varepsilon c(n)}{\alpha}. \end{aligned}$$

We now want the value of  $i$  for which  $|S^{T-i}| > n/2 + \frac{\varepsilon c(n)}{\alpha} > n/2$ . In other words, we want a value of  $i$  such that

$$\beta c(n)(1+\alpha)^i - \frac{\varepsilon c(n)(1+\alpha)^{i+1}}{\alpha} + \frac{\varepsilon c(n)}{\alpha} > n/2 + \frac{\varepsilon c(n)}{\alpha},$$

which is obtained when  $i = T_1$ . Therefore, it is easy to see that if we set  $T = 2T_1$ , we get  $|S^{T_1}| > n/2$ , thereby completing the proof.  $\square$

At first glance, it might appear to be counterintuitive that the order of the bound  $T$  decreases with increasing churn. When the adversary has the benefit of churn that is linear in  $n$ , our bound on  $T$  is a constant, but when the adversary is limited to a churn order of  $\sqrt{n}$ , we get  $T \in O(\log n)$ . This, however, turns out to be fairly natural when we note that the size of the set  $U$  of nodes that we start out with is in proportion to the churn limit.

We say that a node  $u \in V^r$  is *suppressed for  $R$  rounds* if  $|\text{INFLUENCE}_r(u, R)| < n - \beta c(n)$ ; otherwise we say it is *unsuppressed*. The following lemma shows that, given a set with cardinality at least  $\beta c(n)$ , some node in that set will be unsuppressed.

**Lemma 2.** *Consider the adaptive adversary. Let  $U$  be any subset of  $V^{r-1}$ ,  $r \geq 1$ , such that  $|U| \geq \beta c(n)$ . Let  $T$  be the bound derived in Lemma 1. There is at least one  $u^* \in U$  such that for some  $R \in O(T \log n)$ ,*

$$|\text{INFLUENCE}_r(u^*, R)| > n - \beta c(n).$$

*In particular, when the order of the churn is  $n$ ,  $T$  becomes a constant, and we have  $R = O(\log n)$ .*

Before we proceed with our key arguments of the proof, we state a property of bipartite graphs that we will use subsequently.

**Property 1.** Let  $H = (A, B, E)$  be a bipartite graph in which  $|A| > 1$  and every vertex  $b \in B$  has at least one neighbor in  $A$ . There is a subset  $A^* \subset A$  of cardinality at most  $\lceil |A|/2 \rceil$  such that  $|\{b : \exists a^* \in A^* \text{ such that } (a^*, b) \in E\}| \geq \lceil |B|/2 \rceil$ .

*Proof of Property 1.* Consider each node in  $A$  to be a unique color. Color each node in  $B$  using the color of a neighbor in  $A$  chosen arbitrarily. Now partition  $B$  into maximal subsets of nodes with like colors. Consider the parts of the partition sorted in decreasing order of their cardinalities. We now greedily choose the first  $\lceil |A|/2 \rceil$  colors in the sorted order of parts of  $B$ . We call the chosen colors  $C$ . Clearly, colors in  $C$  cover at least as many nodes in  $B$  as those not in  $C$ . Suppose the colors in  $C$  cover fewer than  $\lceil |B|/2 \rceil$  nodes in  $B$ . Then the remaining colors will cover  $\lceil |B|/2 \rceil$ , but that is a contradiction. Therefore, colors in  $C$  cover at least  $\lceil |B|/2 \rceil$  nodes in  $B$ . The nodes in  $A$  that have the colors in  $C$  are the nodes that comprise  $A^*$ , thereby completing our proof.  $\square$

*Proof of Lemma 2.* Again, our proof assumes  $r = 1$  because it generalizes to arbitrary values of  $r$  quite easily. From Lemma 1, we know that the influence of all nodes in  $U$  taken together will reach  $n - \beta c(n)$  nodes in  $T$  rounds. This does not suffice because we are interested in showing that there is at least one node in  $V^0$  that (individually) influences  $n - \beta c(n)$  nodes in  $V^R$  for some  $R = O(T \log n)$ .

From Lemma 1, we know that  $U$  (collectively) will influence at least  $n - \beta c(n)$  nodes in  $T$  rounds, i.e.,

$$|\text{INFLUENCE}(U, T)| > n - \beta c(n).$$

From Property 1, we know that there is a set  $U_1 \subset U$  of cardinality at most  $\lceil |U|/2 \rceil$  such that

$$|\text{INFLUENCE}(U_1, T)| > \frac{n - \beta c(n)}{2}.$$

Recalling that  $\beta < \frac{1}{12} < \frac{1}{3}$ , we know that  $|\text{INFLUENCE}(U_1, T)| \geq \beta c(n)$ . We can again use Lemma 1 to say that  $\text{INFLUENCE}(U_1, T)$  influences more than  $n - \beta c(n)$  nodes in additional  $T$  rounds and, by transitivity,  $U_1$  influences more than  $n - \beta c(n)$  nodes after  $2T$  rounds. We therefore have  $|\text{INFLUENCE}(U_1, 2T)| > n - \beta c(n)$ . Again, we can choose a set  $U_2 \subset U_1$  (using Property 1) that consists of  $\lceil |U_1|/2 \rceil$  nodes in  $U_1$  such that  $|\text{INFLUENCE}(U_2, 2T)| \geq \beta c(n)$ . Subsequently applying Lemma 1 extends the influence set of  $U_2$  to more than  $n - \beta c(n)$  after  $3T$  rounds.

In every iteration  $i$  of the above argument, the size of the set  $U_i$  decreases by a constant fraction until we are left with a single node  $u^* \in U$  such that  $|\text{INFLUENCE}(u^*, O(\log n)T)| > n - \beta c(n)$ .  $\square$

Can  $\beta c(n)$  (or more nodes) be suppressed for any significant number of (say,  $\Omega(T \log n)$ ) rounds? This is in immediate contradiction to Lemma 2 because any such suppressed set of nodes must contain an unsuppressed node! This leads us to the following corollary.

**Corollary 1.** *The number of nodes that can be suppressed for  $\Omega(T \log n)$  rounds is less than  $\beta c(n)$ , even if the networks is controlled by an adaptive adversary.*

**Corollary 2.** *Consider an oblivious adversary that must commit to the entire sequence of graphs in advance. If we choose a node  $u$  uniformly at random from  $V^0$ , with probability at least  $1 - \frac{\beta c(n)}{n}$ ,*

$$|\text{INFLUENCE}(u, \Omega(T \log n))| > n - \beta c(n).$$

*Proof.* Let  $S \subset V^0$  be the set of nodes suppressed for  $\Omega(T \log n)$  rounds. Under an oblivious adversary, the node  $u$  chosen uniformly at random from  $V^0$  will not be in  $S$  with probability  $1 - \frac{\beta c(n)}{n}$ , and hence, will not be suppressed with that same probability.  $\square$

**Lemma 3.** *Consider a dynamic network under linear churn that is controlled by an adaptive adversary. In some  $O(\log n)$  rounds, there is a set of unsuppressed nodes  $V^* \subseteq V^0$  of cardinality more than  $(1 - \beta)n$  such that*

$$\left| \bigcap_{v \in V^*} \text{INFLUENCE}(v, r) \right| > (1 - \beta)n.$$

*Proof.* Let  $V^* \subseteq V^0$  be any set of unsuppressed nodes, i.e., in some  $c_0 \log n$  rounds for some constant  $c_0$ , the influence set of each  $v \in V^*$  has cardinality more than  $(1 - \beta)n$ . Note, however, that we *cannot* guarantee that, for any two vertices  $v_1$  and  $v_2$  in  $V^*$ ,  $|\text{INFLUENCE}(v_1, c_0 \log n) \cap \text{INFLUENCE}(v_2, c_0 \log n)| > (1 - \beta)n$ . Assume for simplicity that  $|V^*|$  is a power of 2. Consider any pair of vertices  $\{v_1, v_2\}$ , both members of  $V^*$ . Recalling that  $\beta < \frac{1}{12} < \frac{1}{3}$ , we can say that  $|\text{INFLUENCE}(v_1, c_0 \log n) \cap \text{INFLUENCE}(v_2, c_0 \log n)| \geq \beta n$ . Therefore, considering the intersected set  $\text{INFLUENCE}(v_1, c_0 \log n) \cap \text{INFLUENCE}(v_2, c_0 \log n)$  of nodes has cardinality at least  $\beta n$ , we can apply Lemma 1 leading to  $|\text{INFLUENCE}(v_1, c_0 \log n + T) \cap \text{INFLUENCE}(v_2, c_0 \log n + T)| > (1 - \beta)n$ . We can partition  $V^*$  into a set  $S_1$  of  $\frac{|V^*|}{2}$  pairs such that for each pair, the intersection of influence sets has cardinality more than  $(1 - \beta)n$  after  $c_0 \log n + T$  rounds. Similarly, we can construct a set  $S_2$  of quadruples by disjointly pairing the pairs in  $S_1$ . Using similar argument, we can say that for any  $Q \in S_2$ ,  $|\bigcap_{v \in Q} \text{INFLUENCE}(v, c_0 \log n + 2T)| > (1 - \beta)n$ . Progressing similarly, the set  $S_{\log |V^*|}$  will equal  $V^*$  and we can conclude that

$$\left| \bigcap_{v \in S_{\log |V^*|}} \text{INFLUENCE}(v, c_0 \log n + T \log |V^*|) \right| > (1 - \beta)n.$$

Since  $|V^*| \leq n$ ,  $c_0 \log n + T \log |V^*| \in O(\log n)$ , thus completing the proof.  $\square$

**Lemma 4.** *Suppose that up to  $\varepsilon\sqrt{n}$  nodes can be subjected to churn in any round by an adaptive adversary. In some  $r \in O(\log^2 n)$  rounds, there is a set of unsuppressed nodes  $V^* \subseteq V^0$  of cardinality at least  $n - \beta\sqrt{n}$  such that*

$$\left| \bigcap_{v \in V^*} \text{INFLUENCE}(v, r) \right| > n - \beta\sqrt{n}.$$

*Proof.* From Corollary 1, we know that each of the unsuppressed nodes in  $V^*$  (which is of cardinality more than  $n - \beta\sqrt{n}$ ) will influence more than  $n - \beta c(n)$  nodes in  $O(\log^2 n)$  time. We can use the same argument as in Lemma 3 to show that in  $O(\log n)$  rounds, all the unsuppressed nodes have a common influence set of size at least  $\Theta(n)$ . That common influence set will grow to all but  $n - \beta\sqrt{n}$  nodes within another  $O(\log^2 n)$  rounds. Thus a total of  $O(\log^2 n)$  rounds is sufficient to fulfill the requirements.  $\square$

### 3.3 Maintaining Information in the Network

In a dynamic network with churn limit  $\varepsilon n$ , the entire set of nodes in the network can be churned out and new nodes churned in within  $1/\varepsilon$  rounds. How do the new nodes even know what algorithm is running? How do they know how far the algorithm has progressed? To address these basic questions, the network needs to maintain some global information that is not lost as the nodes in the network are churned out. There are two basic pieces of information that need to be maintained so that a new node can join in and participate in the execution of the distributed algorithm:

1. the algorithm that is currently executing, and
2. the number of rounds that have elapsed in the execution of the algorithm. In other words, a global clock has to be maintained.

We assume that the nodes in  $V^0$  are all synchronized in their understanding of what algorithm to execute and the global clock. The nodes in the network continuously flood information on what algorithm is running so that when a new node arrives, unless it is shielded by churn, it receives this information and can start participating in the algorithm. To maintain the clock value, nodes send their current clock value to their immediate neighbors. When a new node receives the clock information from a neighbor, it sets its own clock accordingly. Since nodes are not malicious or faulty, Lemma 1 ensures that information is correctly maintained in more than  $n - \beta c(n)$  nodes.

### 3.4 Support Estimation Under an Oblivious Adversary

Suppose we have a dynamic network with  $\mathcal{R}$  nodes colored red in  $V^0$ .  $\mathcal{R}$  is also called the *support* of red nodes. We want the nodes in the network to estimate  $\mathcal{R}$  under an oblivious adversary. We assume that the adversary chooses  $\mathcal{R}$  and which  $\mathcal{R}$  nodes in  $V^0$  to color red, but it does not know the random choices made by the algorithm. Furthermore, we assume that churn can be linear in  $n$ , i.e.,  $c(n) = n$ .

We now provide our algorithm.  $P \in O(\log n)$  is the number of parallel iterations performed by our algorithm in order to increase the precision of our estimate to hold with high probability. Its exact value is worked out in the proof of Theorem 1. At round 1, each red node in  $V^0$  draws  $P$  random samples  $s_1, s_2, \dots, s_i, \dots, s_P$ , each from the exponential random distribution with rate 1. Each  $s_i$  is chosen with a precision that ensures that the smallest possible positive value is at most  $\frac{1}{n^{\Theta(1)}}$ ; note that  $O(\log n)$  bits suffice. Each red node  $u$  initiates  $P$  parallel flooding messages  $m_u(i)$ ; each  $m_u(i)$  contains  $s_i$  and its terminating condition is: HAS ENCOUNTERED A MESSAGE  $m_v(i)$  WITH A SMALLER RANDOM SAMPLE. Note that for  $i \neq j$ , messages  $m_u(i)$  do not interact with messages  $m_u(j)$ . This allows each live node  $u$  to keep track of the  $P$  smallest samples that it has seen, which

we denote as  $\bar{s}_u(i)$  for each  $i$ . After some  $t \in O(\log n)$  rounds, each node  $u \in V^t$  computes the average  $\bar{s}_u$  over all  $\bar{s}_u(i)$  that it has. Each node  $u$  estimates  $\mathcal{R}$  to be  $1/\bar{s}_u$ . It is easy to see that the number of bits transmitted per round through a link is at most  $O(\log^2 n)$ .

To analyze this algorithm, we use two properties of exponential random variables. Consider  $K \geq 1$  independent random variables  $Y_1, Y_2, \dots, Y_K$ , each following the exponential distribution of rate  $\lambda$ .

**Property 2** (e.g., see [16]). The minimum among all  $Y_i$ ,  $1 \leq i \leq K$ , is an exponentially distributed random variable with parameter  $K\lambda$ .

**Property 3** (see [28] and pp. 30 and 35 of [12]). Let  $X_K = \frac{1}{K} \sum_{i=1}^K Y_i$ . Then, for any  $\varsigma \in (0, 1/2)$ ,

$$\Pr \left( \left| X_K - \frac{1}{\lambda} \right| \geq \frac{\varsigma}{\lambda} \right) \leq 2 \exp \left( -\frac{\varsigma^2 K}{3} \right).$$

**Theorem 1.** *Consider an oblivious adversary. With high probability,  $(1 - \beta)n$  nodes in the network estimate  $\mathcal{R}$*

- *between  $(1 - \delta)\mathcal{R}$  and  $(1 + \delta)\mathcal{R}$  for some arbitrarily small  $\delta > 2\beta$  when  $\mathcal{R}$  is large, say  $\mathcal{R} \geq n/2$ , and*
- *between  $\mathcal{R} - \delta n$  and  $\mathcal{R} + \delta n$  when  $\mathcal{R}$  is small, say  $\mathcal{R} < n/2$ .*

*Proof.* Suppose  $\mathcal{R} \geq n/2$ . Out of the  $\mathcal{R}$  red nodes up to  $\beta n$  nodes (chosen obliviously) can be suppressed leaving us with

$$\mathcal{R}' \geq \mathcal{R} - \beta n \geq (1 - 2\beta)\mathcal{R} \quad (4)$$

unsuppressed red nodes (since  $\mathcal{R} \geq n/2$ ). In a slight abuse of notation, we use  $\mathcal{R}$  and  $\mathcal{R}'$  to denote both the cardinality and the set of red nodes and unsuppressed red nodes, respectively. Let

$$U = \{u : u \in \bigcap_{v \in \mathcal{R}'} \text{INFLUENCE}(v, t)\};$$

Note that  $t = O(\log n)$  and  $|U| \geq (1 - \beta)n$  (cf. Lemma 3). Let  $u$  be some node in  $U$ . Let

$$V_u = \{v : v \in \mathcal{R} \wedge u \in \text{INFLUENCE}(v, t)\}.$$

For all  $u \in U$ ,  $\mathcal{R}' \subseteq V_u \subseteq \mathcal{R}$ . Intuitively, at round  $t$ , node  $u$  is estimating  $\mathcal{R}$  using the exponential random numbers that were drawn by nodes in  $V_u$ . Since our adversary is oblivious, the choice of  $V_u$  is independent of the choice of the random numbers generated by each  $v \in V_u$ . Therefore,  $\bar{s}_u(i)$  is an exponentially distributed random number with rate  $|V_u| \geq \mathcal{R}'$  (cf. Property 2). For any  $\delta > 2\beta$ , let  $\varsigma \leq \min\{\frac{\delta - 2\beta}{1 - \delta}, \frac{\delta}{1 + \delta}\}$ . When  $P = \frac{3c \ln n}{\varsigma^2} \in O(\log n)$  parallel iterations are performed, where  $c > 0$ , the required accuracy is obtained with probability  $1 - \frac{1}{\Omega(n^c)}$  (cf. Property 3). The case where  $\mathcal{R} < n/2$  can be addressed in like manner. However, we need to allow an error range that is dependent on  $n$  as up to  $\beta n$  nodes can be suppressed.  $\square$

### 3.5 Support Estimation Under an Adaptive Adversary

The algorithm for support estimation under an oblivious adversary (cf. Section 3.4) does not work under an adaptive adversary. To estimate the support of red nodes in the network, each red node draws a random number from the exponential distribution and floods it in an attempt to spread the smallest random number. When the adversary is adaptive, the smallest random numbers can easily be targeted and suppressed. To mitigate this difficulty, we consider a different algorithm in which the number of bits communicated is more. In particular, the number of bits communicated per round by each node executing this algorithm is at most polynomial in  $n$ .



Let  $\mathcal{R}$  be the support of the red nodes. Every node floods its unique identifier along with a bit that indicates whether it is a red node or not. At most  $\beta\sqrt{n}$  nodes' identifiers can be suppressed by the adversary for  $\Omega(\log^2 n)$  rounds leaving at least  $n - \beta\sqrt{n}$  unsuppressed identifiers (cf. Corollary 1). Each node counts the number of unique red identifiers  $A$  and non-red identifiers  $B$  that flood over it and estimates  $\mathcal{R}$  to be  $A + \frac{n-A-B}{2}$ .

This support estimation technique generalizes quite easily to arbitrary churn order. Therefore, we state the following theorem more generally.

**Theorem 2.** *Consider the algorithm mentioned above in which nodes flood their unique identifiers indicating whether they are red nodes or not and assume that the network is controlled by an adaptive adversary. Let  $c(n)$  be the order of the churn; we assume for simplicity that  $c(n)$  is either  $n$  or  $\sqrt{n}$ . Then the following holds:*

1. *At least  $n - \beta c(n)$  nodes estimate  $\mathcal{R}$  between  $\mathcal{R} - \frac{\beta c(n)}{2}$  and  $\mathcal{R} + \frac{\beta c(n)}{2}$ . Furthermore, these nodes are aware that their estimate is within  $\mathcal{R} - \frac{\beta c(n)}{2}$  and  $\mathcal{R} + \frac{\beta c(n)}{2}$ .*
2. *The remaining nodes are aware that their estimate of  $\mathcal{R}$  might fall outside  $[\mathcal{R} - \frac{\beta c(n)}{2}, \mathcal{R} + \frac{\beta c(n)}{2}]$ . When  $c(n) = n$ , it requires only  $O(\log n)$  rounds, but when  $c(n) = \sqrt{n}$ , it requires  $O(\log^2 n)$  rounds.*

*Proof.* Let  $u$  be any one of the  $n - \beta c(n)$  nodes that receive at least  $n - \beta c(n)$  unsuppressed identifiers (cf. Lemma 3 and Lemma 4). Let  $A$  and  $B$  be the number of unique identifiers from red nodes and non-red nodes, respectively, that flood over  $u$ . Let  $C = n - A - B \leq \beta c(n)$ . This means that  $u$  estimates  $\mathcal{R}$  to be  $A + \frac{C}{2}$ . Clearly,  $A \leq \mathcal{R} \leq A + C$  and since  $C \leq \beta c(n)$ ,  $\mathcal{R}$  is estimated between  $\mathcal{R} - \frac{\beta c(n)}{2}$  and  $\mathcal{R} + \frac{\beta c(n)}{2}$ . Furthermore, since  $u$  received  $n - \beta c(n)$  identifiers, it can be sure that its estimate is between  $\mathcal{R} - \frac{\beta c(n)}{2}$  and  $\mathcal{R} + \frac{\beta c(n)}{2}$ .

If a node does not receive at least  $n - \beta c(n)$  identifiers, then it is aware that its estimate of  $\mathcal{R}$  might not be within  $[\mathcal{R} - \frac{\beta c(n)}{2}, \mathcal{R} + \frac{\beta c(n)}{2}]$ .

From Lemma 3, when  $c(n) = n$ , the algorithm takes  $O(\log n)$  rounds to complete because we want to ensure that unsuppressed nodes have flooded the network. When  $c(n) = \sqrt{n}$ , as a consequence of Lemma 4, the algorithm requires  $O(\log^2 n)$  rounds.  $\square$

## 4 STABLE AGREEMENT Under an Oblivious Adversary

In this section we will first present Algorithm 1 for the simpler problem of reaching BINARY CONSENSUS, where the input values are restricted to  $\{0, 1\}$  (cf. Section 2.1). We will then use this algorithm as a subroutine for solving STABLE AGREEMENT in Section 4.2.

Throughout this section we assume suitable choices of  $\varepsilon$  and  $\alpha$  such that the upper bound

$$\beta < \frac{1}{12} \tag{5}$$

can be satisfied for  $\beta$ ; note that (5) must hold in addition to bound (1) on page 5. Moreover, we assume that a node can send an process up to  $O(\log^2 m)$  bits in every round, where  $m$  is the size of the input value domain.

### 4.1 BINARY CONSENSUS

A node  $u$  that executes Algorithm 1 proceeds in a sequence of  $O(\log n)$  checkpoints that are interleaved by  $O(\log n)$  rounds. Each node  $u$  has a bit variable  $b_u$  that stores its current output value. At each checkpoint  $t_i$ , node  $u$  initiates support estimation of the number of nodes currently having 1 as output bit by using the algorithm described in Section 3.4. (At checkpoint  $t_{R-1}$ , nodes estimate both: the support of 1 and 0.) The outcome of this support estimation will be available in

checkpoint  $t_{i+1}$  where  $u$  has derived the estimation  $\#(1)$ . If  $u$  believes that the support of 1 is small ( $\leq \frac{1}{4}n$ ), it sets its own output  $b_u$  to 0; if, on the other hand,  $\#(1)$  is large ( $\geq \frac{3}{4}n$ ),  $u$  sets its output  $b_u$  to 1. This guarantees stability once agreement has been reached by a large number of nodes. When the support of 1 is roughly the same as the support of 0, we need a way to sway the decision to one side or the other. This is done by flooding the network whereby the flooding messages are weighted by some randomly chosen value. The adversary can only guess which node has the highest weight and therefore, with constant probability, the flooding message with this highest weight (i.e., smallest random number) will be used to set the output bit by almost all nodes in the network.

---

**Algorithm 1** BINARY CONSENSUS under an oblivious adversary; code executed by node  $u$ .

---

Let  $decision_u$  be initialized to  $\perp$ .

Let  $b_u$  be the current output bit of  $u$ . Initially, for each  $u \in V^0$ ,  $b_u$  is set to the input value assigned to  $u$ .

Let  $t_1 = 1$  be the first checkpoint round. Subsequent checkpoint rounds are given by  $t_i = t_{i-1} + O(\log n)$ , for  $i > 1$ . Node  $u$  decides at round  $t_R$ , for some  $R = O(\log n)$ , thereby requiring  $O(\log^2 n)$  rounds.

**At every checkpoint round  $t_i$  including  $t_1$ :**

- 1: Initiate support estimation (to be completed in checkpoint round  $t_{i+1}$ ).
- 2: Generate a random number  $r_u$  uniformly from  $\{1, \dots, n^k\}$  for suitably large but constant  $k$ . (With high probability, we want exactly one node to have generated  $\min_u r_u$ .)
- 3: Initiate flooding of  $\{r_u, b_u\}$  with terminating condition:  $\langle (\text{HAS ENCOUNTERED ANOTHER MESSAGE INITIATED BY } v \neq u \text{ WITH } r_v < r_u) \vee (\text{CURRENT ROUND} \geq t_{i+1}) \rangle$ .

**At every checkpoint round  $t_i$  except  $t_1$ :**

- 4: Use the support estimation initiated at checkpoint round  $t_{i-1}$ . Let  $\#(1)$  be  $u$ 's estimated support value for the number of nodes that had an output of 1.
- 5: **if**  $\#(1) \leq \frac{1}{4}n$  **then**
- 6:    $b_u \leftarrow 0$ .
- 7: **else if**  $\#(1) \geq \frac{3}{4}n$  **then**
- 8:    $b_u \leftarrow 1$ .
- 9: **else if**  $u$  has received flooded messages initiated in  $t_{i-1}$  **then**
- 10:   Let  $\{r_v, b_v\}$  be the message with the smallest random number that flooded over  $u$ .
- 11:    $b_u \leftarrow b_v$ .

**At terminating checkpoint round  $t_R$ :**

- 12: **if**  $\#(1) \geq \frac{n}{2}$  **then**
- 13:    $decision_u \leftarrow 1$ .
- 14:   Flood a 1-decision message ad infinitum.
- 15: **else if**  $\#(0) \geq \frac{n}{2}$  **then**
- 16:    $decision_u \leftarrow 0$ .
- 17:   Flood 0-decision message ad infinitum.

**If  $u$  receives a  $b$ -decision message:**

- 18:  $decision_u \leftarrow b$
- 

**Theorem 3.** Assume that the adversary is oblivious and that the churn limit per round is  $\epsilon n$ . Algorithm 1 solves BINARY CONSENSUS in  $O(\log^2 n)$  rounds with high probability.

*Proof.* We first argue that Validity holds: Suppose that all nodes start with input value 1. The only way a node can set its output to 0 is by passing Line 5. This can happen for at most  $\beta n$  nodes. The only way that more nodes can set their output to 0 is if they estimate the support of 1 to be in  $(\frac{1}{4}n, \frac{3}{4}n)$ . If  $\beta$  is suitably small, Theorem 1 guarantees that with high probability this will not happen at any node. The argument is analogous for the case where all nodes start with 0.

Next we show Almost Everywhere Agreement: Let  $N_i$  be the number of nodes at checkpoint round  $t_i$  that output 1. Let  $LO_i$ ,  $HI_i$ , and  $MID_i$ , respectively, be the sets of nodes in  $V^{t_i}$  for which  $\#(1) \leq \frac{1}{4}n$ ,  $\#(1) \geq \frac{3}{4}n$ , and  $\frac{1}{4}n < \#(1) < \frac{3}{4}n$ ; note that nodes are placed in  $LO_i$ ,  $HI_i$ ,

and  $\text{MID}_i$  based on their  $\#(1)$  values, which are estimates of  $N_{i-1}$ , not  $N_i$ . Clearly, we have that  $\text{LO}_i + \text{MID}_i + \text{HI}_i = n$ .

For some  $i > 1$ , let  $u^* \in V^{t_{i-1}}$  be the node that generated the smallest random number in checkpoint round  $t_{i-1}$  among all nodes in  $V^{t_{i-1}}$ . With high probability,  $u^*$  will be unique. By Corollary 2, with probability  $1 - \beta$  (a constant),  $u^*$  is unsuppressed, implying that  $b_{u^*}$  will be used by all nodes in  $\text{MID}_i$ . Consider the following cases:

**Case A** ( $N_{i-1} \leq (\frac{1}{4} - \delta)n$ ): From Theorem 1, we know that with high probability  $|\text{LO}_i| \geq (1 - \beta)n$  implying  $|\text{MID}_i| + |\text{HI}_i| \leq \beta n$ . Therefore,  $N_i$  will continue to be very small leading to small estimates  $\#(1)$  in subsequent checkpoints. After  $O(\log n)$  rounds, this causes at least  $(1 - \beta)n$  nodes to decide on 0, with high probability. Moreover, it is easy to see that the remaining  $\beta n$  nodes will not be able to pass Line 12, since the adversary cannot artificially increase the estimated support of nodes with 0. (Recall from Section 3.4 that by suppressing the minimum random variables, the adversary can only make the estimate smaller.)

**Case B** ( $(\frac{1}{4} - \delta)n < N_{i-1} < (\frac{1}{4} + \delta)n$ ): With high probability,  $|\text{LO}_i| + |\text{MID}_i| \geq (1 - \beta)n$  implying  $|\text{HI}_i| \leq \beta n$ . Note first that nodes in  $\text{LO}_i$  will set their output bits to 0. Since  $N_{i-1} < (\frac{1}{4} + \delta)n$ , there are at least  $(\frac{3}{4} - \varepsilon)n$  nodes in  $V^{t-1}$  that output 0. Of these, at most  $\beta n$  could have been suppressed. So, with probability at least  $\frac{3}{4} - \delta - \beta$ ,  $u^*$  is an unsuppressed nodes that outputs 0. When  $u^*$  outputs 0, nodes in  $\text{MID}_i$  will set their output bits to 0. Thus, considering  $\text{LO}_i$  and  $\text{MID}_i$ , we have at least  $(1 - \beta)n$  nodes that set their output bits to 0 with constant probability. For a suitably small  $\delta$  and  $\beta < \frac{1}{4} - \delta$ , this will lead to Case A in the next iteration, which means that subsequently nodes agree on 0.

**Case C** ( $(\frac{1}{4} + \delta)n \leq N_{i-1} \leq (\frac{3}{4} - \delta)n$ ): With high probability,  $|\text{MID}_i| \geq (1 - \beta)n$ . With constant probability  $(1 - \beta)$ ,  $u^*$  will be an unsuppressed node and nodes in  $\text{MID}_i$  will set their output bits to the same value  $b_{u^*}$ .

**Case D** ( $(\frac{3}{4} - \delta)n < N_{i-1} < (\frac{3}{4} + \delta)n$ ): This is similar to Case B, i.e., with constant probability, at least  $(1 - \beta)n$  nodes will reach agreement on 1.

**Case E** ( $N_{i-1} \geq (\frac{3}{4} + \delta)n$ ): This is similar to Case A. With high probability, at least  $(1 - \beta)n$  nodes will decide on 1.

Note that, when a checkpoint falls either under Case A or Case E, with high probability, it will remain in that case. When a checkpoint falls under Case B, Case C, or Case D, with constant probability, we get either Case A or Case E in the following checkpoint. Therefore, in  $O(\log n)$  rounds, at least  $(1 - \beta)n$  nodes will reach agreement with high probability and the all other nodes will remain undecided.

For property Stability, note that if a node has decided on some value in checkpoint  $t_R$ , it continues to flood its decision message. We showed that, with high probability, at least  $(1 - \beta)n$  nodes will decide on the same bit value. Therefore, it follows by Lemma 1 that agreement will be maintained ad infinitum among at least  $(1 - \beta)n$  nodes.  $\square$

In order to use Algorithm 1 to solve STABLE AGREEMENT, we will need to make a couple of crucial adaptations.

- Suppose every vertex in  $V^0$  has some auxiliary information. We can easily adapt Algorithm 1 so that when a node  $u$  decides on a bit value  $b$ , then, it also inherits the auxiliary information of some  $v \in V^0$  whose initial bit value was  $b$ . This adaptation is possible because our algorithm ensures Validity.

- For a typical agreement algorithm, we assume that all nodes simultaneously start running the algorithm consensus. We want to adapt our algorithm so that only nodes in  $V^0$  that have an initial output bit of 1 initiate the algorithm, while nodes that start with 0 are considered passive, i.e., these nodes do not generate messages themselves, but still forward flooding messages and start generating messages from the next checkpoint onward as soon as they notice that an instance of the algorithm is running.

We now sketch how the algorithm can be adapted: In the first checkpoint  $t_1$ , each node  $v$  with a 1 initiates support estimation and flooding of message  $\langle r_v, b_v = 1 \rangle$ . If the number of nodes with 1 is small at checkpoint  $t_1$ , then, at checkpoint  $t_2$ , nodes that receive estimate values will conclude 0, which will get reinforced in subsequent checkpoints. However, if the number of nodes with a 1 at checkpoint  $t_1$  is large (in particular, larger than  $\beta n$ ), then, by suitable flooding most nodes (in particular, at least  $(1 - \beta)n$  nodes) will know that a support estimation is underway and will participate from checkpoint  $t_2$  onward.

## 4.2 A 3-phase Algorithm for STABLE AGREEMENT

We will now describe how we use Algorithm 1 as a building block for solving STABLE AGREEMENT:

**Flooding Phase:** In the very first round, each node  $u \in V^0$  generates a uniform random number  $r_u$  from  $(0, 1)$  and if the random number is less than  $\frac{\log n}{n}$ , it initiates a message  $m_u$  for flooding. The message  $m_u$  contains the random number  $r_u$  and the value  $\text{VAL}_u$  assigned to  $u$  by the adversary. Nodes enter the candidate selection phase (see below) after a sufficient number of rounds to ensure that no more than  $\beta n$  nodes are suppressed (see Corollary 1). However, the flooding messages go on ad infinitum.

**Candidates Selection Phase:** We initiate an expected  $O(\log n)$  parallel iterations of BINARY CONSENSUS, each associated with one of the (expected)  $O(\log n)$  flooding messages  $m_u$ . More precisely, the instance of BINARY CONSENSUS for a particular  $m_u$  is designed as follows: nodes that have received the flooded message  $m_u$  set themselves to 1 and initiate BINARY CONSENSUS. If  $m_u$  has reached saturation (i.e., flooded to at least  $(1 - \beta)n$  nodes), the consensus value will be 1. If  $m_u$  has a very small support (say,  $\beta n$ ), the consensus value will be 0 with high probability (cf. Case A of the proof of Theorem 3). When the support of  $m_u$  is neither too small nor too large, the nodes will reach consensus on either 0 or 1. We say that a flooded message  $m_u$  is a *candidate* message if the instance of BINARY CONSENSUS associated with it reached a consensus value 1. Note that, with high probability,  $(1 - \beta)n$  nodes agree on the set of candidate messages. Among the candidate messages, every node  $v$  chooses the message  $m_u$  with the smallest random number  $r_u$  and value  $\text{VAL}_u$ , and initiates a support estimation for  $m_u$ .

**Confirmation Phase:** On expectation,  $\log n$  nodes initiate flooding in the Flooding phase. From Corollary 2, each of them will not be suppressed with probability at least  $(1 - \beta)$ . Therefore, with high probability, at least one node  $u$  will have  $|\text{INFLUENCE}(u, O(\log n))| \geq (1 - \beta)n$ . That is, at least one flooded message  $m_v$  will be a candidate message and therefore, when the support estimation is initiated, a set  $S$  of at least  $(1 - \beta)n$  nodes will measure its support to be at least  $(1 - \beta - \delta)n$  for some  $\delta > 2\beta$  with high probability (cf. Theorem 1). Due to (5), there can only be one such message  $m_v$  with high support. The nodes  $S$  will decide on the attached  $\text{VAL}_v$  of  $m_v$ , whereas nodes that do not observe that  $m_v$  has high support (because of being shielded by churn nodes) remain undecided. This shows almost everywhere agreement.

Analogously to Algorithm 1, nodes in  $S$  flood their decision messages, which are adopted by newly incoming nodes. By virtue of Lemma 1, the stability property is maintained ad infinitum.

The additional running time overhead of the above three phases excluding Algorithm 1 is only in  $O(\log n)$ . Thus we have shown the following result:

**Theorem 4.** *Consider the oblivious adversary and suppose that  $\varepsilon n$  nodes can be subject to churn in every round. The 3-phase algorithm is correct with high probability and reaches STABLE AGREEMENT in  $O(\log^2 n)$  rounds.*

## 5 STABLE AGREEMENT Under an Adaptive Adversary

In this section we consider the STABLE AGREEMENT problem while dealing with a more powerful adaptive adversary. At the beginning of a round  $r$ , this adversary observes the entire state of the network and previous communication between nodes (including even previous outcomes of random choices!), and thus can adapt its choice of  $G^r$ , to make it much more difficult for nodes to achieve agreement.

It is instructive to consider the algorithms presented in Section 4 in this context. Both approaches are doomed to fail in the presence of an adaptive adversary: For the STABLE AGREEMENT algorithm, the expected number of nodes that initiate flooding in the flooding phase is  $\log n$ . Even though each of these nodes would have expanded its influence set to some constant size by the end of the next round, the adaptive adversary can spot and immediately churn out all these nodes before they can communicate with anyone else, thus none of these values will gain any support. Simply increasing the order of the expected number of flooding nodes to match the churn limit does not help, as this will cause considerable amount of congestion and therefore slow down the spreading rate of the flooding; this in turn will cause the runtime of the algorithm to exceed  $O(\log n)$ .

Algorithm 1 fails for the simple reason that the adversary can selectively suppress the flooding of the smallest generated random value  $z \in \{1, \dots, n^k\}$  with attached bit  $b_z$  from ever reaching some 50% of the nodes, which instead might use a distinct minimum value  $z'$  (with an attached bit value  $b_{z'} \neq b_z$ ) to guide their output changes.

To counter the difficulties we have mentioned, we relax the model. Firstly, we limit the order of the churn to  $\sqrt{n}$ . Secondly, we allow messages of up to  $O(n)$  bits to be sent over a link in a single round. Under these relaxations, we can estimate the support of red nodes in the network simply by flooding all the unique identifiers of the red and non-red nodes (cf. Theorem 2).

Similarly to Section 4, we will first solve BINARY CONSENSUS under these assumptions and then show how to implement STABLE AGREEMENT. In this section we assume that the number of nodes in the network is sufficiently large, such that

$$n \gg 4\beta^2. \quad (6)$$

Moreover, we assume that every node can send and process up to  $O(n + \log m)$  bits per round, where  $m$  is the size of the input domain.

### 5.1 BINARY CONSENSUS

We now describe an algorithm for solving BINARY CONSENSUS, which is similar in spirit to Algorithm 1. The main difference is the handling of the case where the support of the nodes that output 1 is roughly equal to the support of the nodes with output bit 0. In this case we rely on the variance of random choices made by individual nodes to sway the balance of the support towards one of the two sides with constant probability.



---

**Algorithm 2** BINARY CONSENSUS under an adaptive adversary; code executed by node  $u$ .

---

Let  $decision_u$  be initialized to  $\perp$ .

Let  $b_u$  be the current output bit of  $u$ . Initially, for each  $u \in V^0$ ,  $b_u$  is set to the input value assigned to  $u$ .

Let  $t_1 = 1$  be the first checkpoint round. Subsequent checkpoint rounds are given by  $t_i = t_{i-1} + O(\log^2 n)$ ,  $i > 1$ , with time between consecutive checkpoint rounds sufficient for unsuppressed nodes to reach a common influence(cf. Lemma 4).

The algorithm terminates at round  $t_R$ , for some  $R = O(\log n)$ , thereby requiring  $O(\log^3 n)$  rounds.

**At every checkpoint round  $t_i$  including  $t_1$ , but excluding  $t_R$ :**

- 1: Initiate support estimation (to be completed in checkpoint round  $t_{i+1}$ ).

**At every checkpoint round  $t_i$  excluding  $t_1$  and  $t_R$ :**

- 2: Use the support estimation initiated at checkpoint round  $t_{i-1}$ . Let  $\#(1)$  be the estimated support value for nodes that output 1.
- 3: **if** support estimation is not accurate within  $[\mathcal{R} - \frac{\beta\sqrt{n}}{2}, \mathcal{R} + \frac{\beta\sqrt{n}}{2}]$  **then**
- 4:   Do nothing.
- 5: **else if**  $\#(1) < \frac{n}{2} - \frac{\beta\sqrt{n}}{2}$  **then**
- 6:    $b_u \leftarrow 0$ .
- 7: **else if**  $\#(1) > \frac{n}{2} + \frac{\beta\sqrt{n}}{2}$  **then**
- 8:    $b_u \leftarrow 1$ .
- 9: **else**
- 10:   **if** the outcome of an unbiased coin flip is heads **then**
- 11:      $b_u \leftarrow 0$ .
- 12:   **else**
- 13:      $b_u \leftarrow 1$ .

**At terminating checkpoint round  $t_R$ :**

- 14: **if**  $\#(1) \geq \frac{n}{2}$  **then**
- 15:    $decision_u \leftarrow 1$ .
- 16:   Flood a 1-decision message ad infinitum.
- 17: **else if**  $\#(0) \geq \frac{n}{2}$  **then**
- 18:    $decision_u \leftarrow 0$ .
- 19:   Flood a 0-decision message ad infinitum.

**If  $u$  receives a  $b$ -decision message:**

- 20:  $decision_u \leftarrow b$
- 

**Theorem 5.** *Algorithm 2 solves BINARY CONSENSUS within  $O(\log^3 n)$  rounds with high probability, even in the presence of an adaptive adversary and up to  $\varepsilon\sqrt{n}$  churn per round.*

*Proof.* First consider property Validity: Suppose that all nodes start with input value 1. Theorem 2 guarantees that any node  $u$  that receives insufficient many identifiers for support estimation, will execute Line 4 and therefore never set its output to 0. On the other hand, if  $u$  does receive sufficiently many samples, again Theorem 2 ensures that it will always pass the if-check in Line 7. Thus, no node can every output 0. The case where all nodes start with 0 can be argued analogously.

Next, we will show that Algorithm 2 satisfies almost everywhere agreement. Let  $N_i$  be the number of vertices at checkpoint round  $t_i$  that output 1. Let  $LO_i$ ,  $HI_i$ , and  $MID_i$ , respectively, be the sets of nodes in  $V^{t_i}$  for which  $\#(1) \leq n/2 - \frac{\beta\sqrt{n}}{2}$ ,  $\#(1) \geq n/2 + \frac{\beta\sqrt{n}}{2}$ , and  $n/2 - \frac{\beta\sqrt{n}}{2} < \#(1) < n/2 + \frac{\beta\sqrt{n}}{2}$ ; note that nodes are placed in  $LO_i$ ,  $HI_i$ , and  $MID_i$  based on their  $\#(1)$  values, which are estimates of  $N_{i-1}$ , not  $N_i$ . In a slight abuse of notation, we use  $LO_i$ ,  $MID_i$ , and  $HI_i$  to also refer to their respective cardinalities. Clearly, we have that  $LO_i + MID_i + HI_i = n$ . Furthermore, observe that either  $LO_i$  or  $HI_i$  will be 0. Otherwise, we will have two nodes such that one estimates  $N_{i-1}$  below  $n/2 - \frac{\beta\sqrt{n}}{2}$ , while the other estimates it above  $n/2 + \frac{\beta\sqrt{n}}{2}$  — a violation of Theorem 2.

Consider the following cases.

**Case A** ( $N_{i-1} < n/2 - \beta\sqrt{n}$ ): From Theorem 2,  $LO_i \geq n - \beta\sqrt{n}$  and all nodes in  $LO_i$  will set themselves to output 0. Once this case is reached in some checkpoint, it will be reached in all future checkpoints until  $t_R$  with high probability. Therefore, the algorithm guarantees almost everywhere agreement on 0 in  $t_R$ ; with high probability, nodes do not pass Line 14 in checkpoint  $t_R$ , thus no node will ever decide on 1.

**Case B** ( $N_{i-1} > n/2 + \beta\sqrt{n}$ ): This case is similar to Case A with the difference that almost all nodes decide on 1.

**Case C** ( $n/2 - \beta\sqrt{n} \leq N_{i-1} \leq n/2$ ): Notice that  $HI_i = 0$ . Therefore,

$$LO_i + MID_i \geq n - \beta\sqrt{n}. \quad (7)$$

We consider two subcases:

1. In this case, we assume that  $LO_i$  is at least  $n/2 + \beta\sqrt{n}$ . This will set  $N_i < n/2 - \beta\sqrt{n}$  putting the network in Case A in the next checkpoint.
2. In this case, we assume that  $LO_i < n/2 + \beta\sqrt{n}$ . This implies that  $MID_i \geq n - LO_i - \beta\sqrt{n} \geq n/2 - 2\beta\sqrt{n}$ . The nodes in  $MID_i$  will choose 1 or 0 with equal probability. The number of nodes that choose 0 is a binomial distribution with mean  $\frac{MID_i}{2}$  and standard deviation  $\frac{\sqrt{MID_i}}{2}$ . Clearly, with some constant probability,  $\frac{MID_i}{2} + \frac{\sqrt{MID_i}}{2}$  or more nodes in the set  $MID_i$  will set themselves to output 0. Therefore, with constant probability,

$$\begin{aligned} N_i &< n - LO_i - \frac{MID_i}{2} - \frac{\sqrt{MID_i}}{2} \\ &< n - LO_i - \frac{n - LO_i - \beta\sqrt{n}}{2} - \frac{\sqrt{n - LO_i - \beta\sqrt{n}}}{2} \end{aligned}$$

Clearly,  $N_i < \frac{n}{2} - \beta\sqrt{n}$  if

$$\begin{aligned} 3\beta\sqrt{n} &< \sqrt{n - LO_i - \beta\sqrt{n}}, \\ \Rightarrow 9\beta^2 n &< n - LO_i - \beta\sqrt{n}, \\ \Rightarrow LO_i + \beta\sqrt{n} &< n - 9\beta^2 n. \end{aligned}$$

We know that  $LO_i < \frac{n}{2} + \beta\sqrt{n}$ . Therefore,  $N_i < \frac{n}{2} - \beta\sqrt{n}$  if

$$\begin{aligned} \frac{n}{2} + 2\beta\sqrt{n} &< n - 9\beta^2 n, \\ \Rightarrow 2\beta\sqrt{n} &< \frac{n}{2} - 9\beta^2 n, \\ \Rightarrow 2\beta &< \sqrt{n} \left( \frac{1}{2} - 9\beta^2 \right). \end{aligned}$$

In other words, as long as

$$n > \frac{4\beta^2}{\left(\frac{1}{2} - 9\beta^2\right)^2}, \quad (8)$$

with constant probability,  $N_i < \frac{n}{2} - \beta\sqrt{n}$ , which will put the network in Case A at the next checkpoint round. The bound (6) guarantees that Condition (8) is easily met.

**Case D** ( $n/2 < N_{i-1} \leq n/2 + \beta\sqrt{n}$ ): Using arguments similar to Case C, we can show that with constant probability,  $N_i > \frac{n}{2} + \beta\sqrt{n}$ , thereby, putting the network in Case B.

Clearly, after  $O(\log n)$  checkpoint rounds, with high probability, the network will reach either Case A or Case B<sup>3</sup> and hence achieve almost everywhere agreement on either 0 or 1.

For property Stability, note that if a node has decided on some value  $\neq \perp$  in checkpoint  $t_R$ , it continues to flood its decision message. Since at least  $(1 - \beta)n$  have decided, it follows by Lemma 1 that any nodes that have been churned in will also decide on this value within a constant number of rounds, thus agreement will be maintained ad infinitum.  $\square$

## 5.2 STABLE AGREEMENT

Now that we have a solution for BINARY CONSENSUS, we will show how to use it to solve STABLE AGREEMENT where nodes have input values from some set  $\{0, \dots, m\}$ , for  $m \geq 1$ . Given some input value  $\text{VAL}$  we can write it in the base-2 number system as  $(b_0, \dots, b_{\log m})$  where  $b_i \in \{0, 1\}$ , for  $1 \leq i \leq \log m$ . We call  $\text{VAL}$  a *general input value* and  $b_i$  a *binary input value*.

The basic idea of the STABLE AGREEMENT algorithm is to run an instance of the BINARY CONSENSUS algorithm for each  $b_i$  and then combine the agreed bits to obtain agreement on the general input values. More specifically, in the first instance every node uses the bit  $b_0$  of its general input value as binary input for the BINARY CONSENSUS algorithm. We need to be careful, however, to not violate the validity property of STABLE AGREEMENT. Thus we assume that every node sends its general input value along with the input bit. When the BINARY CONSENSUS instance of node  $u$  decides on some bit value  $b$ , node  $u$  overwrites its general input value with the input value  $\text{VAL}_b$  that was sent along with  $b$ . For the next instance of BINARY CONSENSUS,  $u$  uses the second bit of  $\text{VAL}_b$  and so on. After  $\log m$  such instances, we can be sure that the sequence of binary decision values corresponds to the bit value of some general input value, thus guaranteeing validity. Stability and Almost Everywhere Agreement follow from the properties of BINARY CONSENSUS.

**Theorem 6.** *Suppose that the network is controlled by an adaptive adversary who can subject up to  $\varepsilon\sqrt{n}$  nodes to churn in every round. There is an algorithm that solves STABLE AGREEMENT in  $O(\log m \log^3 n)$ .*

## 6 Impossibility of a Deterministic Solution

In this section we show that there is no deterministic algorithm to solve STABLE AGREEMENT even when the churn is restricted to only a constant number of nodes per round. As a consequence, randomization is a necessity for solving STABLE AGREEMENT.

We introduce some well known standard notations (see [3, Chap. 5]) used for showing impossibility results of agreement problems. The *configuration*  $C^r$  of the network at round  $r$  consists of

- the graph of the network at that point in time, and
- the local state of each node in the network.

A specific run  $\rho$  of some STABLE AGREEMENT algorithm  $\mathcal{A}$  is entirely determined by an infinite sequence of configurations  $C^0, C^1, \dots$  where  $C^0$  contains the initial state of the graph before the first round. Consider the input value domain  $\{0, 1\}$ . A configuration  $C^r$  is *1-valent* (resp., *0-valent*) if all possible runs of  $\mathcal{A}$  that share the common prefix up to and including  $C^r$ , lead to an agreement value of 1 (resp., 0). Note that this decision value refers to the decision of the large majority of nodes; strictly speaking, a small fraction of nodes might remain undecided on  $\perp$ . A configuration

---

<sup>3</sup>Due to Equation (6) we know that Cases A and B exist.

is *univalent* if it is either 1-valent or 0-valent. Any configuration that is not univalent is called a *bivalent* configuration.

**Lemma 5.** *Consider a bivalent configuration  $C^r$  in round  $r$  reached by an algorithm  $\mathcal{A}$  that solves STABLE AGREEMENT and ensures Almost Everywhere Agreement. No node in  $V^r$  can have decided on a value  $\neq \perp$  by round  $r$ .*

*Proof.* Assume in contradiction that some node  $u$  has already decided on 0 in some bivalent configuration  $C^r$ . Then, by the Almost Everywhere Agreement property, no other node  $v$  can ever decide on 1 in the same run. But this means that  $C^r$  is actually a univalent configuration, yielding a contradiction.  $\square$

**Theorem 7.** *Suppose that the sequence of graphs  $(G^r)_{r \geq 0}$  is an expander family with degree  $\Delta$ . Assume that the churn is limited to at most  $\Delta+1$  nodes per round. There is no deterministic algorithm that solves STABLE AGREEMENT if the network is controlled by an adaptive adversary.*

*Proof.* We use an argument that is similar to the argument used in the proof that  $f+1$  rounds are required for consensus in the presence of  $f$  faults (cf. [3, Chap. 5]). For the purpose of this impossibility proof, we restrict the input domain of nodes to  $\{0, 1\}$  and allow arbitrary congestion on the communication channels. Moreover, we assume that the topology of the network is fixed throughout the run. Thus the adversary can only “replace” nodes at the same position by some other nodes.

For the sake of contradiction, assume that such a deterministic algorithm  $\mathcal{A}$  exists that solves STABLE AGREEMENT under the assumed settings. We will prove our theorem by inductively constructing an infinite run  $\rho$  of this algorithm consisting of a sequence of bivalent configurations. By virtue of Lemma 5 this allows us to conclude that nodes do not reach almost everywhere agreement.

To establish the basis of our induction, we need to show that there is an initial bivalent configuration  $C^0$  at the start of round 1. Assume in contradiction that there is no bivalent starting configuration. Clearly, if all nodes start with a value 0 (resp., 1), this network must reach STABLE AGREEMENT on 0 (resp., 1). This implies that there are two possible starting configurations  $C_0^0$  and  $C_1^0$  in which (i) the input values are the same for all but one node  $u^0$ , but (ii)  $C_0^0$  is 0-valent whereas  $C_1^0$  is 1-valent. Consider the respective one-round extension of  $C_0^0$  and  $C_1^0$  where the adversary simply churns out node  $u^0$ . Both successor configurations  $C_0^1$  and  $C_1^1$  are indistinguishable for all other nodes, in particular they have no way of knowing what initial value was assigned to  $u^0$ , since all witnesses have been removed by the adversary. Therefore,  $C_0^1$  and  $C_1^1$  must both be either 0-valent or 1-valent, a contradiction. This shows that there is an initial bivalent configuration, thereby establishing the basis for our induction.

For the inductive step, we assume that the network is in a bivalent configuration  $C^{r-1}$  at the end of round  $r-1$ . We will extend  $C^{r-1}$  by one round (guided by the adversary) that yields another bivalent configuration  $C^r$ . Assume for the sake of a contradiction that every possible one-round extension of  $C^{r-1}$  yields a univalent configuration. Without loss of generality, assume that the one-round extension  $\gamma$  where no node is churned out is 1-valent and yields configuration  $C_1^r$ . Since by assumption  $C^{r-1}$  was bivalent, there is another one-round extension  $\gamma'$  that yields a 0-valent configuration  $C_0^r$ . Moreover, we know that a nonempty set  $S$  of size at most  $\Delta+1$  nodes must have been subject to churn in  $\gamma'$ . (This is the only difference between  $C_0^r$  and  $C_1^r$  — recall that the edges of the graph are stable throughout the run.)

Let  $S'$  be a subset of  $S$  and let  $\gamma_{S'}$  be the one-round extension of  $C^{r-1}$  that we get when only nodes in  $S'$  are churned out. Clearly,  $\gamma = \gamma_{\emptyset}$  and  $\gamma' = \gamma_S$ . Consider the lattice of all such one-round extension bounded by  $\gamma$  and  $\gamma'$  that is given by the power set of  $S$ . Starting at  $\gamma$  and moving towards  $\gamma'$  along some path, we must reach a one-round extension  $\gamma_{\{v_1, \dots, v_k\}}$  that yields a 1-valent

configuration  $D_1^r$ , whereas the next point on this path is some one-round extension  $\gamma_{\{v_1, \dots, v_{k+1}\}}$  that ends in a 0-valent configuration  $D_0^r$ . The only difference between these two extensions is that node  $v_{k+1}$  is churned out in the latter but not in the former extension. Now consider the one-round extensions of  $D_0^r$  and  $D_1^r$  where  $v_{k+1}$  and all its neighbors are churned out, yielding  $D_0^{r+1}$  and  $D_1^{r+1}$ . For all other nodes,  $D_0^r$  and  $D_1^r$  are indistinguishable and therefore they must either both be 0-valent or both be 1-valent. This, however, is a contradiction.  $\square$

Considering that expander graphs usually are assumed to have constant degree, Theorem 7 implies that even if we limit the churn to a constant, the adaptive adversary can still beat any deterministic algorithm.

## 7 Conclusion

We have introduced a novel framework for analyzing highly dynamic distributed systems with churn. We believe that our model captures the core characteristics of such systems: a large amount of churn per round and a constantly changing network topology. Future work involves extending our model to include Byzantine nodes and corrupted communication channels. Furthermore, our work raises some key questions: How much churn can we tolerate in an adaptive setting? Are there algorithms that tolerate linear (in  $n$ ) churn in an adaptive setting? We show that we can tolerate  $O(\sqrt{n})$  churn in an adaptive setting, but it takes a polynomial (in  $n$ ) number of communication bits per round. An intriguing problem is to reduce the number of bits to polylogarithmic in  $n$ .

While the main focus of this paper was achieving agreement among nodes which is one of the most important tasks in a distributed system, we believe that the techniques we have developed are useful building blocks for tackling other tasks like aggregation or leader election in this setting.

## References

- [1] Cloudmark website. <http://cloudmark.com/>.
- [2] James Aspnes, Navin Rustagi, and Jared Saia. Worm versus alert: Who wins in a battle for control of a large-scale network? In *OPODIS*, pages 443–456, 2007.
- [3] Hagit Attiya and Jennifer Welch. *Distributed Computing: Fundamentals, Simulations and Advanced Topics (2nd edition)*. John Wiley Interscience, March 2004.
- [4] Baruch Awerbuch and Christian Scheideler. Group spreading: A protocol for provably secure distributed name service. In *ICALP*, pages 183–195, 2004.
- [5] Amitabha Bagchi, Ankur Bhargava, Amitabh Chaudhary, David Eppstein, and Christian Scheideler. The effect of faults on network expansion. *Theory Comput. Syst.*, 39(6):903–928, 2006.
- [6] Hervé Baumann, Pierluigi Crescenzi, and Pierre Fraigniaud. Parsimonious flooding in dynamic graphs. In *PODC*, pages 260–269, 2009.
- [7] Piotr Berman and Juan A. Garay. Fast consensus in networks of bounded degree. *Distributed Computing*, 7(2):67–73, 1993.
- [8] John F. Canny. Collaborative filtering with privacy. In *IEEE Symposium on Security and Privacy*, pages 45–57, 2002.



- [9] Keren Censor Hillel and Hadas Shachnai. Partial information spreading with application to distributed maximum coverage. In *Proceeding of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing*, PODC '10, pages 161–170. ACM, 2010.
- [10] Edith Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. Syst. Sci.*, 55(3):441–453, 1997.
- [11] Souptik Datta, Kanishka Bhaduri, Chris Giannella, Ran Wolff, and Hillol Kargupta. Distributed data mining in peer-to-peer networks. *IEEE Internet Computing*, 10(4):18–26, 2006.
- [12] A. Dembo and O. Zeitouni. Large deviations techniques and applications. *Elearn*, 1998.
- [13] Benjamin Doerr, Leslie Ann Goldberg, Lorenz Minder, Thomas Sauerwald, and Christian Scheideler. Stabilizing consensus with the power of two choices. In *SPAA*, pages 149–158, 2011.
- [14] Cynthia Dwork, David Peleg, Nicholas Pippenger, and Eli Upfal. Fault tolerance in networks of bounded degree. *SIAM J. Comput.*, 17(5):975–988, 1988.
- [15] Amos Fiat and Jared Saia. Censorship resistant peer-to-peer content addressable networks. In *SODA*, pages 94–103, 2002.
- [16] C.M. Grinstead and J.L. Snell. *Introduction to probability*. American Mathematical Society, 1997.
- [17] P. Krishna Gummadi, Stefan Saroiu, and Steven D. Gribble. A measurement study of napster and gnutella as examples of peer-to-peer file sharing systems. *Computer Communication Review*, 32(1):82, 2002.
- [18] Kirsten Hildrum and John Kubiawicz. Asymptotically efficient approaches to fault-tolerance in peer-to-peer networks. In *DISC*, volume 2848 of *Lecture Notes in Computer Science*, pages 321–336. Springer, 2003.
- [19] Bruce M. Kapron, David Kempe, Valerie King, Jared Saia, and Vishal Sanwalani. Fast asynchronous byzantine agreement and leader election with full information. *ACM Transactions on Algorithms*, 6(4), 2010.
- [20] Valerie King and Jared Saia. Breaking the  $O(n^2)$  bit barrier: scalable Byzantine agreement with an adaptive adversary. In *PODC*, pages 420–429, 2010.
- [21] Valerie King, Jared Saia, Vishal Sanwalani, and Erik Vee. Scalable leader election. In *SODA*, pages 990–999, 2006.
- [22] Valerie King, Jared Saia, Vishal Sanwalani, and Erik Vee. Towards secure and scalable computation in peer-to-peer networks. In *FOCS*, pages 87–98, 2006.
- [23] F. Kuhn and R. Oshman. Dynamic networks: Models and algorithms. *SIGACT News*, 42(1):82–96, 2011.
- [24] Fabian Kuhn, Rotem Oshman, and Yoram Moses. Coordinated consensus in dynamic networks. In *PODC*, pages 1–10, 2011.
- [25] C. Law and K.-Y. Siu. Distributed construction of random expander networks. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 3, pages 2133 – 2143 vol.3, march-3 april 2003.

- [26] Nancy Lynch. *Distributed Algorithms*. Morgan Kaufman Publishers, Inc., San Francisco, USA, 1996.
- [27] David J. Malan and Michael D. Smith. Host-based detection of worms through peer-to-peer cooperation. In Vijay Atluri and Angelos D. Keromytis, editors, *WORM*, pages 72–80. ACM Press, 2005.
- [28] Damon Mosk-Aoyama and Devavrat Shah. Fast distributed algorithms for computing separable functions. *IEEE Transactions on Information Theory*, 54(7):2997–3007, 2008.
- [29] Moni Naor and Udi Wieder. A simple fault tolerant distributed hash table. In *IPTPS*, pages 88–97, 2003.
- [30] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Building low-diameter p2p networks. In *FOCS*, pages 492–499, 2001.
- [31] Christian Scheideler. How to spread adversarial nodes?: rotate! In *STOC*, pages 704–713, 2005.
- [32] Subhabrata Sen and Jia Wang. Analyzing peer-to-peer traffic across large networks. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement*, IMW '02, pages 137–150, New York, NY, USA, 2002. ACM.
- [33] Daniel Stutzbach and Reza Rejaie. Understanding churn in peer-to-peer networks. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, IMC '06, pages 189–202, New York, NY, USA, 2006. ACM.
- [34] Eli Upfal. Tolerating a linear number of faults in networks of bounded degree. *Inf. Comput.*, 115(2):312–320, 1994.
- [35] Vasileios Vlachos, Stephanos Androutsellis-Theotokis, and Diomidis Spinellis. Security applications of peer-to-peer networks. *Comput. Netw.*, 45:195–205, June 2004.